杭州電子科技大學 HANGZHOU DIANZI UNIVERSITY

WAYNE STATE UNIVERSITY

Bench 2018@Seattle

Power Characterization of Memory Intensive Applications: Analysis and Implications

Yeliang Qiu[1,2], **Congfeng Jiang[1,2]**, Yumei Wang[1,2], Huayou Si[1,2], Bing Luo[3], Weisong Shi[3]

1 Key Laboratory of Complex System Modeling and Simulation, Ministry of Education, Hangzhou 310037, China
2 School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310037, China
3 Wayne State University, Detroit, MI 48201, USA

Outline

- Motivation
- Server Energy Efficiency Evaluation
- Key Findings
- Experiments and Analysis
- Conclusions

Motivation

- In big data paradigm, the processor centric computing is transforming to memory-centric computing.
- Energy efficiency is important for all kinds of servers
- Applications performance is highly correlated with memory capacity and bandwidth.
- Although there are SPECpower and other benchmarks for server performance evaluation, the results can't be a reliable source for energy efficiency evaluation of large memory systems.

Energy Efficiency Evaluation

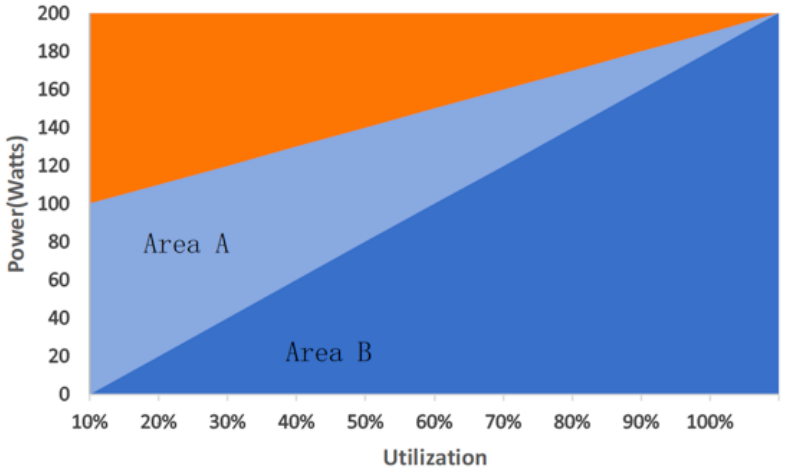
- AN *EXAMPLE OF SPECPOWER SSJ2008 TESTING RESULT IN 2016*

Performance			Power	Performance to Power Ratio
Target Load	Actual Load	ssj_ops	Average Active Power (W)	
100%	99.80%	24,662,648	3,868	6,377
90%	90.10%	22,252,836	3,481	6,393
80%	80.00%	19,758,684	3,032	6,517
70%	70.00%	17,284,975	2,611	6,619
60%	60.00%	14,824,481	2,340	6,336
50%	50.00%	12,350,615	2,143	5,764
40%	40.00%	9,877,126	1,971	5,011
30%	30.00%	7,410,001	1,823	4,064
20%	20.00%	4,949,964	1,674	2,956
10%	10.00%	2,475,968	1,531	1,618
Active Idle		0	1,080	0
			$\sum$ssj_ops/$\sum$power	5,316

Metrics

- $MEE = BpW = \frac{PerceivedBandwid(MB/s)}{SystemPower(watts)}$
- $EE = \frac{Performance}{Power}$
- $EP = 1 - \frac{Area_{real} - Area_{ideal}}{Area_{ideal}}$

$Area_{ideal}$=Area B
$Area_{real}$=Area B+Area A



Key Findings

- Server power consumption changes with workload intensity and concurrent execution threads. However, fully utilized memory systems are not the most energy efficient.
- The memory capacity per processor core has significant impact on the application's performance and server power consumption.
- Memory utilization is not always a good indicator for server power consumption when it's running memory intensive applications.

Experiments and Analysis

- Testbed: Single-node Server

No	*Name*	*Hardware Availability Year*	*CPU Model*	*Total cores*	*CPU TDP (watts)*	*Memory (GB)*	*DISK*
#1	Sugon A620r-G	2012	2*AMD Opteron 6272	32	115	64(8G*8) DDR3 1600MHz	4*SAS 300GB 10K rpm (RAID10)
#2	ThinkServer RD640	2014	2*Intel Xeon E5-2620 #2	12	80	160(16G*10) DDR4 2133MHz	1*SSD 480GB
#3	ThinkServer RD450	2015	2*Intel Xeon E5-2620 #3	12	85	192(16G*12) DDR4 2133MHz	1*SSD 480GB

- Workloads:STREAM,NAMD,CloudSuite and SPECpower

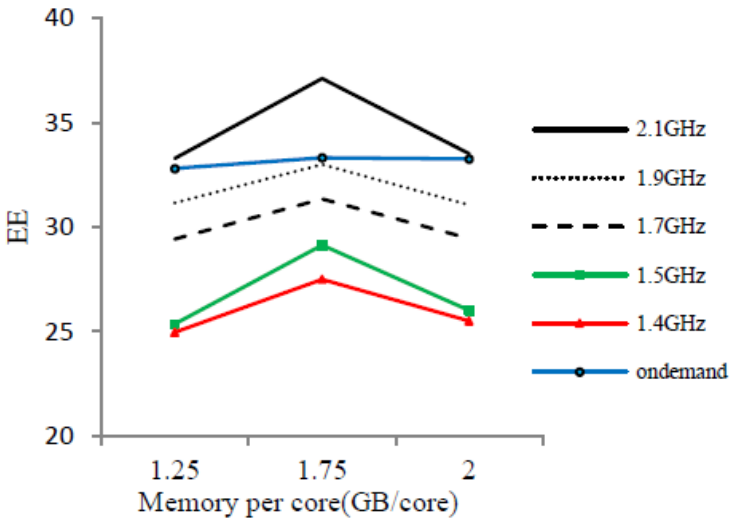
Experiments and Analysis

Configurations:

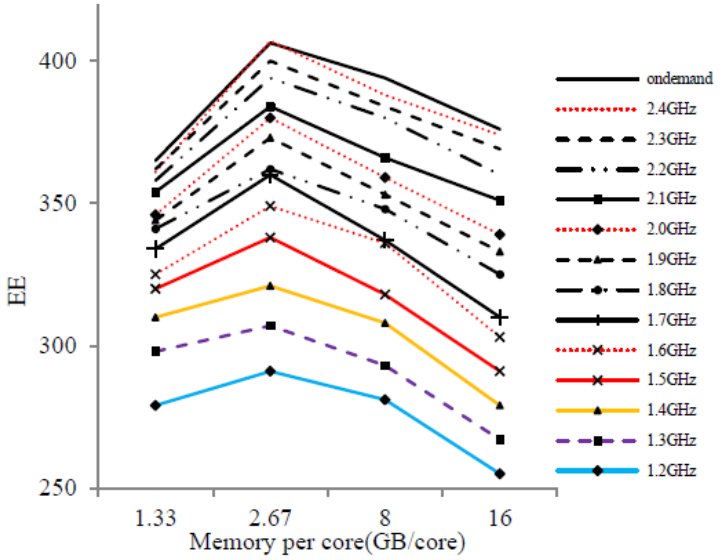
- SPECpower: Default Configuration
- STREAM: Different numbers of concurrent STREAM threads with varying array size
- NAMD: stmv.8M and stmv.28M
- CloudSuite: Default Configuration

Experiments and Analysis

- Results of SPECpower workload



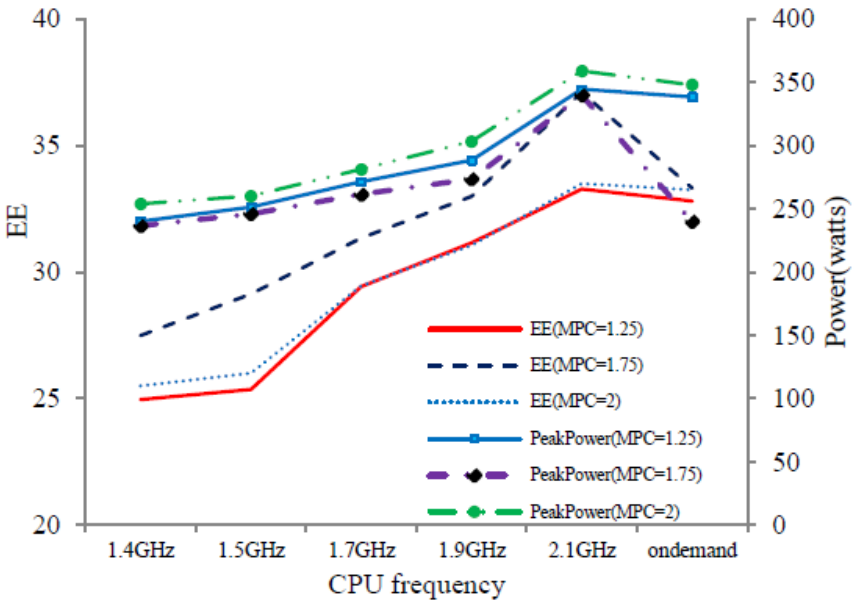
Energy efficiency with different memory per core and CPU frequency on #1 server



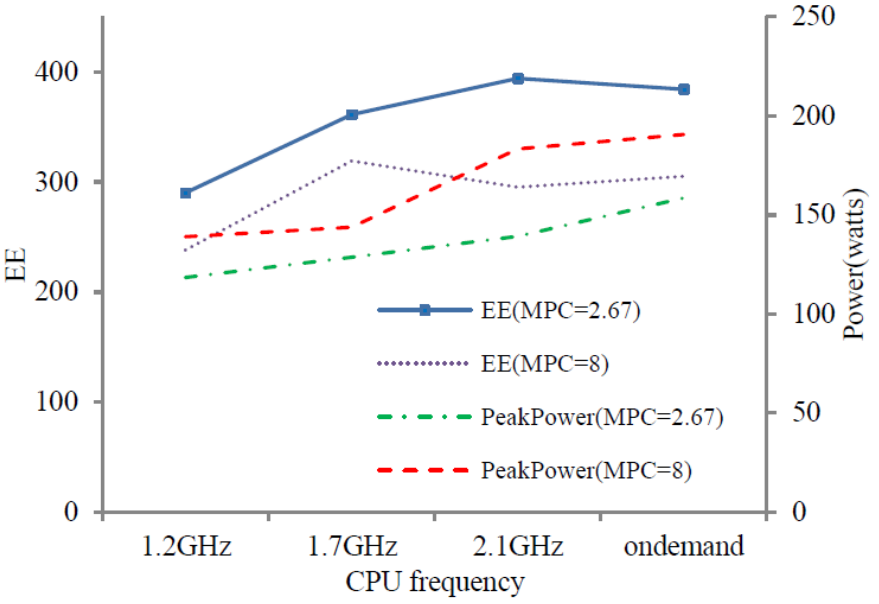
Energy efficiency with different memory per core and CPU frequency on #3 server

Experiments and Analysis

- Results of SPECpower workload



Energy efficiency and peak power on #1 with different memory per core and frequencies



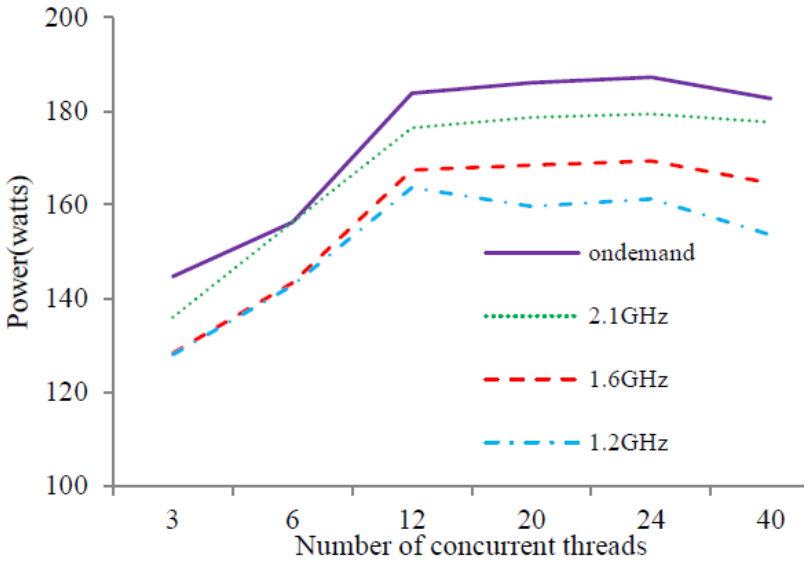
Energy efficiency and peak power on #3 with different memory per core and frequencies

Implication #1

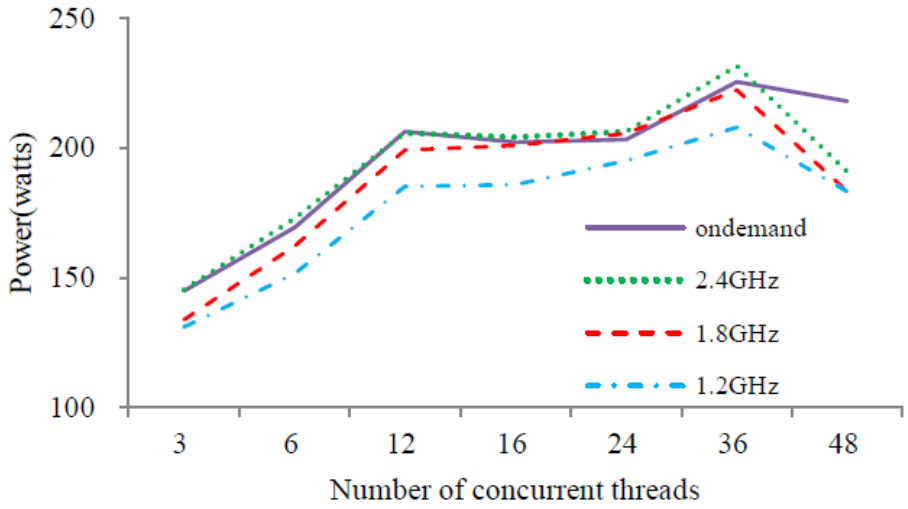
- The on-demand governor always almost has the highest energy efficiency and it's very close to the energy efficiency with the highest frequency.
- Server consumes more power at higher CPU frequency at same memory per core configuration.
- When memory per core configuration increases at fixed CPU frequency, the peak power consumption also increases,but energy efficiency is not.

Experiments and Analysis

- Results of STREAM workload



Power Consumption of server #2 with array size=4GB

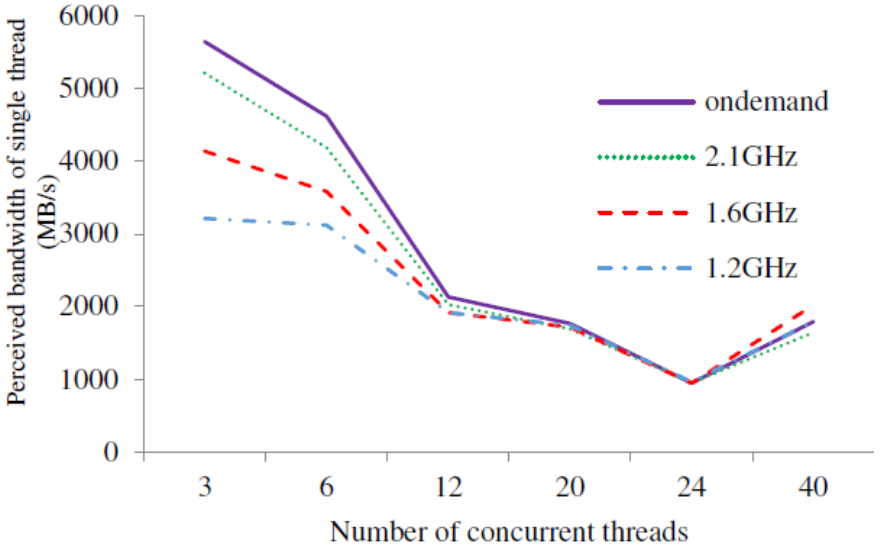


Power Consumption of server#3 with array size=4GB

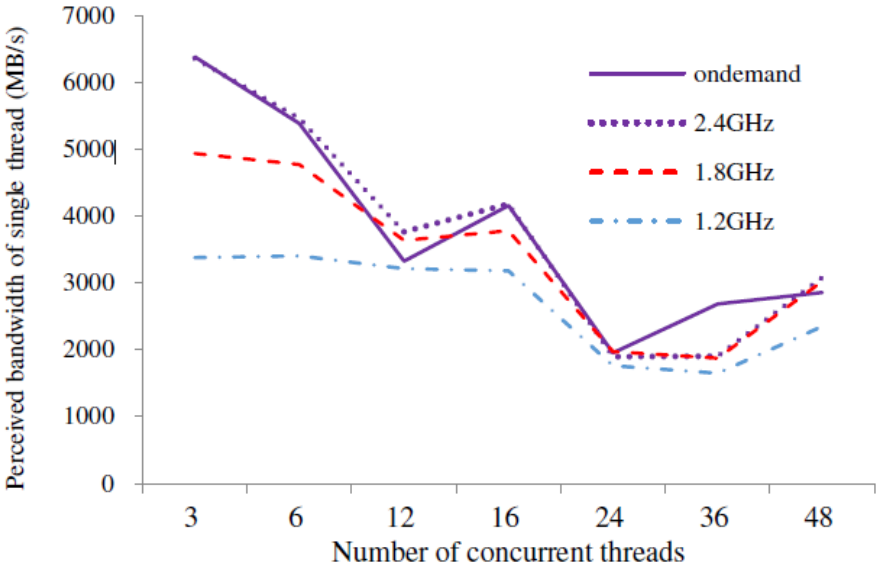
Power increases and then decreases

Experiments and Analysis

- Results of STREAM workload



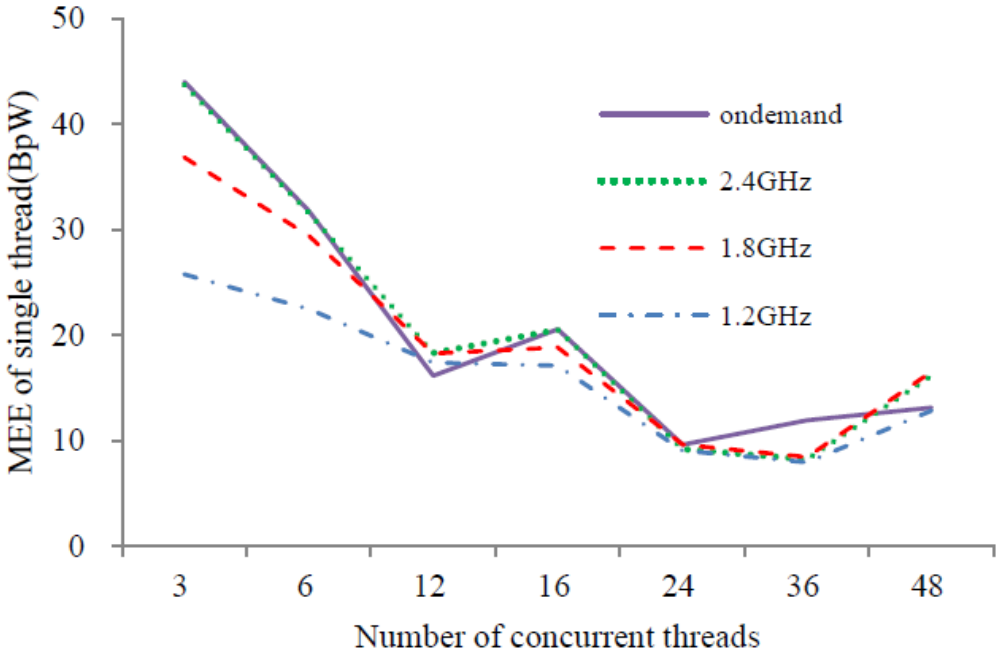
Average Perceived bandwidth of single thread with array size=4GB on server#2



Average Perceived bandwidth of single thread with array size=4GB on server#3

Experiments and Analysis

- Results of STREAM workload



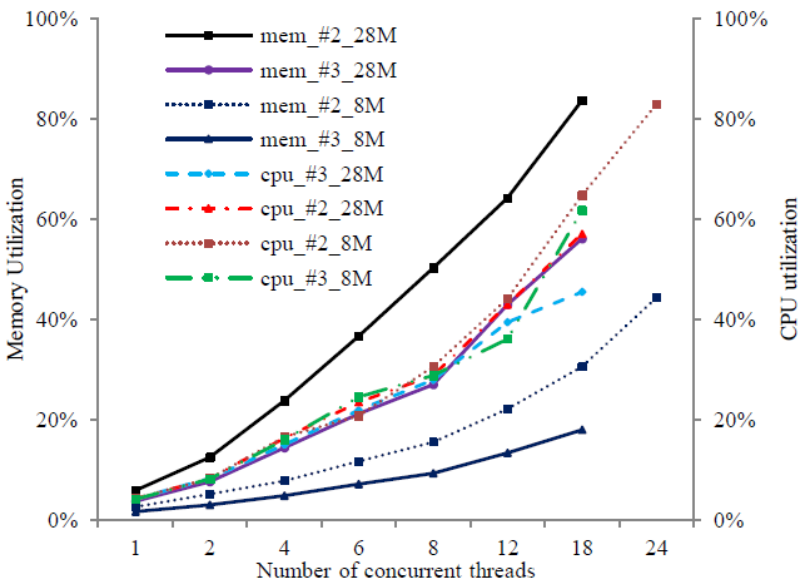
Average memory energy efficiency of single thread with array size=4GB on server#3

Implication #2

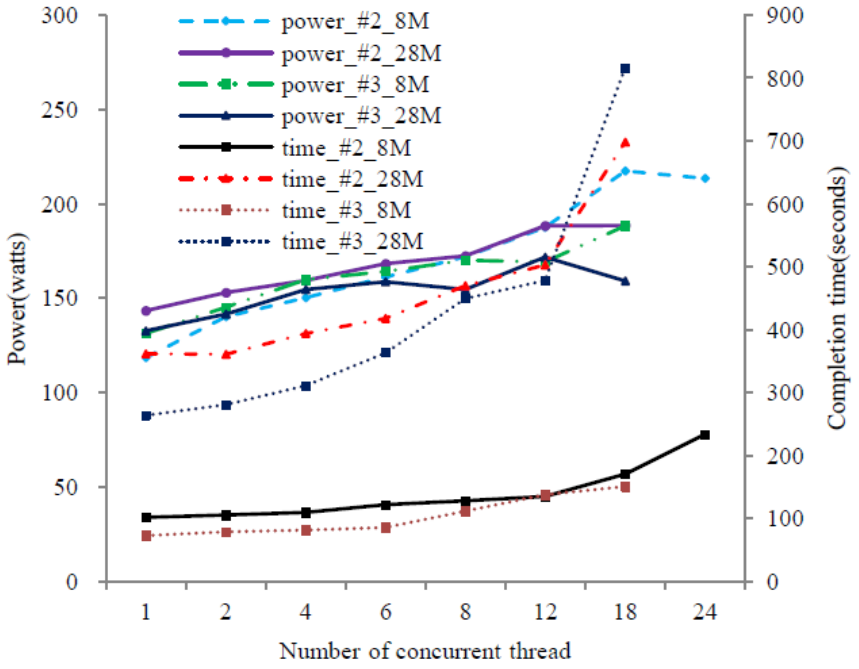
- With increment of concurrent threads and therefore memory utilization, the power consumption of the server also increases.
- The memory energy efficiency decreases when the number of concurrent threads increases.
- And the difference of memory energy efficiency beneficial from CPU frequency with fewer threads is greater than that with more concurrent threads. This means that in a highly contented condition, frequency scaling can’t provide much memory energy efficiency improvements.

Experiments and Analysis

- Results of NAMD workload



CPU and Memory utilization of NAMD



Power and completion time of NAMD

Implication #3

- Results of NAMD workload

Coefficients of correlation between system power and memory

server	power-memory	power-cpu	cpu-memory
#2_8M	0.936	0.958	0.995
#2_28M	0.973	0.966	0.997
#3_8M	0.922	0.938	0.983
#3_28M	0.671	0.750	0.944

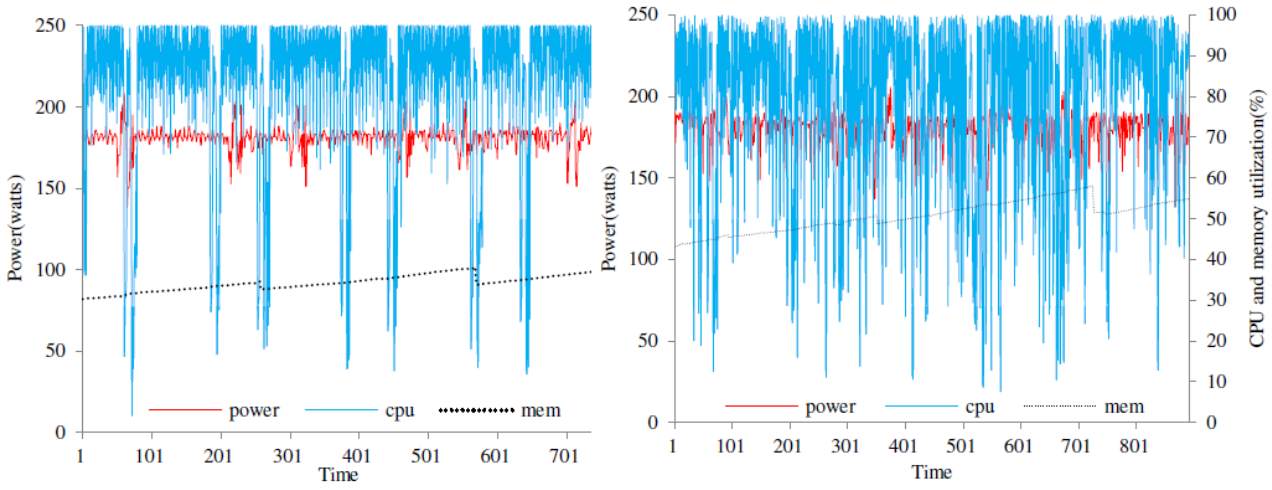
The system power is significantly correlated with the memory and CPU utilization on different machines. Both memory and CPU utilization are good indicators for system power consumption on both server #2 and #3.

Experiments and Analysis

- Results of CloudSuite workload

	power-memory	power-cpu	Memory utilization
IM	-0.57	0.05	0.39
IM_DS	-0.52	-0.09	0.48

Coefficients of correlation between system power and memory

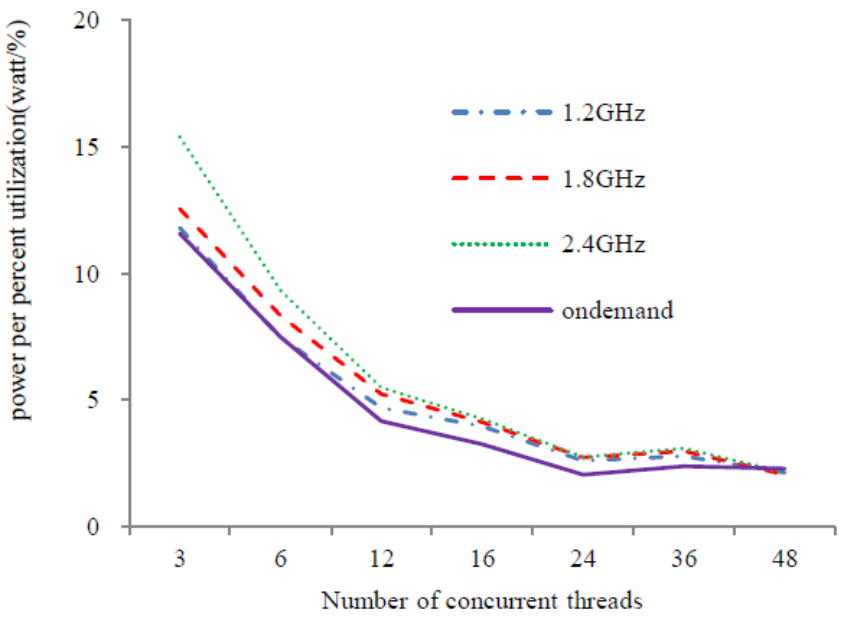


Implication #4

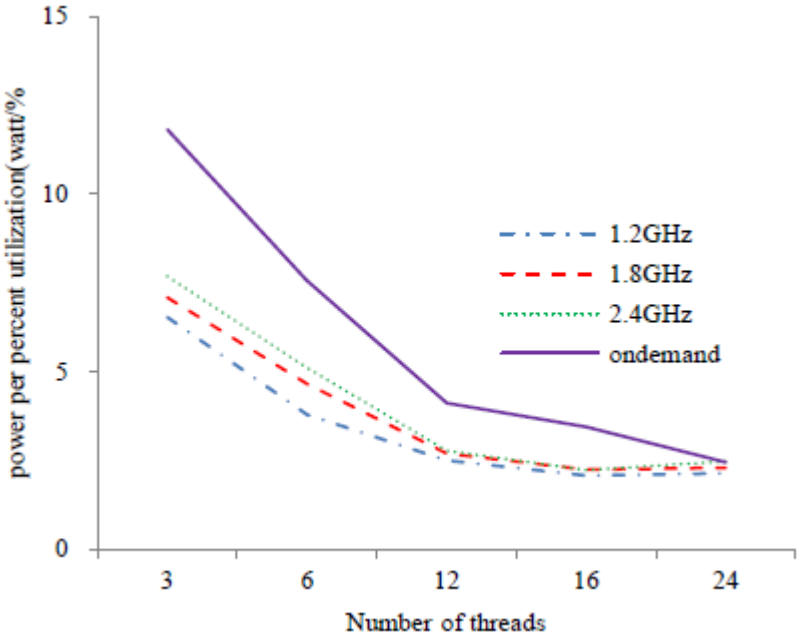
- Neither memory nor CPU utilization is a good indicator for system power consumption.

Experiments and Analysis

- Economies of Scale in Memory Utilization



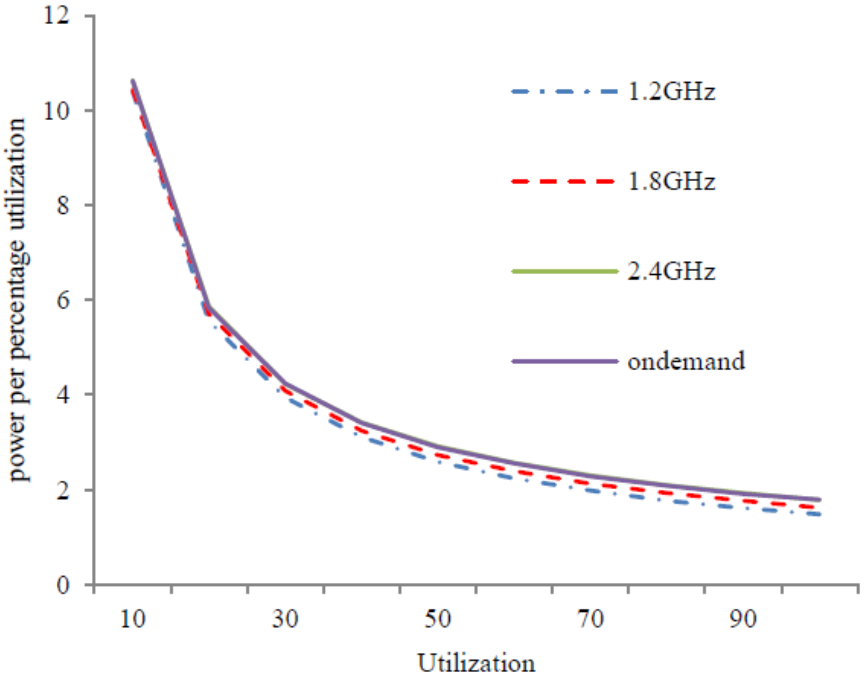
Power per percent memory utilization of server #3 with array size=4GB



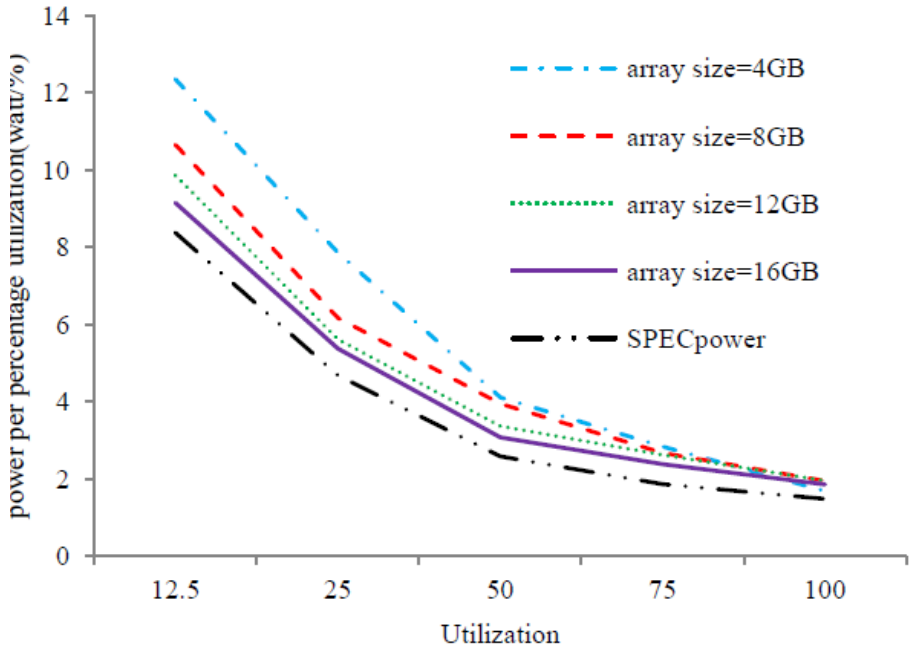
Power per percent memory utilization of server #3 with array size=8GB

Experiments and Analysis

- Economies of Scale in Memory Utilization



Power per percent utilization of server #3 running SPECpower benchmark



Power per percent utilization of #3 running SPECpower and STREAM(CPU frequency at 1.2GHZ)

Implication #5

- When the number of threads increases, the power per percentage memory utilization decreases.

- The power consumption per percent utilization of SPECpower and STREAM benchmark decreases when system utilization increases. However, SPECpower has lower power per percent utilization than STREAM during all utilization levels.

Conclusions

- We conducted extensive experiments and measurements to investigate the power and energy characteristics of three 2U servers running various memory intensive benchmarks.
- Experiment results show that fully utilized memory systems are not the most energy efficient. And the memory capacity per processor core has significant impact on the application's performance and server power consumption
- Memory utilization is not always a good indicator for server power consumption even when it is running memory intensive applications.


杭州電子科技大学
HANGZHOU DIANZI UNIVERSITY

WAYNE STATE
WAYNE STATE
UNIVERSITY

Thanks!